

机器学习的信息科学原理： 基于形式化信息映射的因果链元框架

许建峰

（凯原法学院，智慧司法研究院，计算机学院，上海交通大学，上海，200030，中国）

摘要：

[目的] 聚焦于解决目前机器学习缺乏统一的形式化理论框架、缺乏可解释性和伦理安全保障等问题。

[方法] 本文首先构建形式化信息模型，运用合式公式集合显式定义机器学习各典型环节的本体状态和载体映射，引入可学习和可处理谓词、学习和处理函数分析模型因果链逻辑推演与约束法则。

[结果] 构建了机器学习理论元框架 MLT-MF，以此为基础分别建立了模型可解释性和伦理安全性的普适性定义，证明了模型可解释与信息可还原性、伦理安全保障和泛化误差估计等三个重要定理。

[局限] 当前框架假设理想条件下的信息无噪声使能映射，主要针对静态场景中的模型学习和处理逻辑，同时还未涉及多模态、多智能体系统跨本体空间的信息融合与冲突消解。

[结论] 本文突破碎片化研究局限，为系统解决当前机器学习面临的关键问题提供了统一的理论基础。

关键词：机器学习 客观信息论 形式逻辑系统 可解释性 伦理安全 泛化误差

分类号：TP080717

Information Science Principles of Machine Learning: A Causal Chain Meta-Framework Based on Formalized Information Mapping

Jianfeng Xu

（Koguan School of Law, China Institute for Smart Justice, School of Computer Science,
Shanghai Jiao Tong University, Shanghai, 200030, China）

Abstract:

[Objective] This study focuses on addressing the current lack of a unified formal theoretical framework in machine learning, as well as the deficiencies in interpretability and ethical safety assurance.

[Methods] A formal information model is first constructed, utilizing sets of well-formed formulas to explicitly define the ontological states and carrier mappings of typical components in machine learning. Learnable and processable predicates, along with learning and processing functions, are introduced to analyze the logical deduction and constraint rules of the causal chains within models.

[Results] A meta-framework for machine learning theory (MLT-MF) is established. Based on this framework, universal definitions for model interpretability and ethical safety are

proposed. Furthermore, three key theorems are proved: the equivalence of model interpretability and information recoverability, the assurance of ethical safety, and the estimation of generalization error.

[Limitations] The current framework assumes ideal conditions with noiseless information-enabling mappings and primarily targets model learning and processing logic in static scenarios. It does not yet address information fusion and conflict resolution across ontological spaces in multimodal or multi-agent systems.

[Conclusions] This work overcomes the limitations of fragmented research and provides a unified theoretical foundation for systematically addressing the critical challenges currently faced in machine learning.

Keywords: machine learning objective information theory formal logic systems interpretability ethical and safety assurance generalization error

1 引言

虽然人工智能是当今信息科学乃至整个科学技术中最为活跃的领域，但现有研究仍然存在以下重要不足：一是缺乏覆盖全生命周期的形式化框架[1-5]。现有机器学习理论研究多聚焦于孤立环节：优化理论侧重训练动态性，泛化分析关注测试性能，而初始化策略与部署应用常被视为独立环节。这种割裂导致无法刻画知识从先验注入、数据驱动更新到任务泛化的连贯演化。例如，PAC-Bayes 框架虽连接先验与泛化误差，却忽略训练过程的时间因果性；元学习虽涉及多阶段适配，但未建立符号化状态传递模型。全生命周期建模的缺失，使得模型设计者难以系统分析“参数如何从初始化经训练最终承载任务知识”这一核心问题。二是信息流因果链的显式表达缺失。传统方法常将数据到模型的转化视为黑箱，损失函数梯度驱动参数更新被简化为纯数学优化，忽视先验知识、数据分布与模型架构间的因果依赖。尽管因果推断领域尝试刻画变量间因果关系，但其在机器学习中多限于特征-标签关联分析，未能构建“先验约束→初始化→梯度传播→泛化能力”的完整因果链。例如，对抗训练通过最大损失提升鲁棒性，但其理论分析极少显式关联初始参数约束与最终抗扰性的因果机制，导致策略设计依赖经验试错。三是可解释性工具的数学基础薄弱[11-15]。当前可解释性方法主要通过局部近似或博弈论分配特征重要性，虽具工程实用性，却缺乏严格的数学公理化定义。例如，SHAP 值基于合作博弈的边际贡献计算，但未形式化其与模型信息载体状态的映射关系，导致解释结果可能违背因果性。此外，从参数反推数据特征依赖启发式算法，缺乏“可还原信息”的数学保证。这种理论薄弱性使得解释结果易受对抗攻击干扰，难以应用于医疗、司法等高可信需求场景。

信息科学作为连接数据、知识和系统的核心学科，为机器学习的理论整合提供了新的视角[16, 17]。客观信息论（OIT）通过公理化方法建立信息的六元组模型，通过本体状态表示信息的内在涵义，通过载体状态体现信息的客观表达，通过使能映射说明信息实现具有的数学满射与物理因果双重属性，为人们理解客观世界、尤其是计算机和信息系统中的信息提供了通用普适的基础模型[18-23]。

在此基础上，本文应用递归方法严格定义了高阶逻辑系统，通过合式公式集合的解释建立了状态的形式化表达方式，完全确立了信息定义中状态表达的数学本质，为 OIT 补齐了最后一个理论空白，也为应用 OIT 分析各种信息科学问题提供了更加有力的工具。继而全面应用 OIT 的理论和方法，从先验知识编码、数据驱动训练到模型泛化应用，首次将先验分布、数据分布、参数更新等核心概念统一为形式化状态表达，建立了覆盖机器学习全流程的数理逻辑体系，并以此论证了机器学习中训练信息的形式化建模与实现、模型学习的形式化建模与实现、模型的信息处理与学习优化等典型环节的因果链逻辑推演与约束法则，形成了机器学习理论元框架（Meta-Framework for Machine Learning Theory, MLT-MF）。最后应用 MLT-MF，针对机器学习领域的若干热点问题，分别建立了模型可解释性和伦理安全性的普适性定义，先后证明了模型可解释与信息可还原性、伦理安全保障和泛化误差估计等三个重要定理，为机器学习理论分析和工程实践提供基于统一形式逻辑框架的有力工具。

2 信息的公设与形式化定义

毫无疑问，作为人工智能的一个关键环节，现今的机器学习是计算机、互联网、大数据和云计算等信息技术快速发展和高度集成的成果。机器学习系统无疑是信息系统的一个重要分支。从信息的基本概念入手，能够为机器学习和人工智能的研究提供统一的理论基础。

2.1 信息的公设

人类虽已进入信息时代，但长期以来，信息科学理论分散在信息论、控制论、信号处理、统计学习、计算理论等不同学科，各自独立发展，对于信息的认识依然众说纷纭，缺少统一的公理基础和理论体系，也是导致人工智能缺乏可解释性和可预测性的一个重要因素。对此，[22]提出了关于信息的四个公设：

公设 1（二元主体） 每个信息都有本体和载体两个主体，其中本体具有信息的本质内涵，载体承载信息的客观形态。

公设 2（存在时间） 信息的内涵和形态都有各自的存在时间。

公设 3（状态表达） 信息的本体和载体都有各自的状态表达。

公设 4（使能映射） 信息的本体状态使能映射到载体状态。

这四个公设简明扼要，符合人们对于信息的通常认识，也兼容各个主流信息科学分支的基本概念，其科学意义体现在：

- * 统一性：为长期分散的信息科学理论提供了共同基础；
- * 公理化：使信息的定义和推理可以像数学和物理学一样公理化、形式化；
- * 普适性：便于推广到互联网、大数据、人工智能、认知科学、信息物理系统等前沿领域。

其中“使能映射”是一个新创的概念，它的涵义是既要求在数学上能够建立本体状态向载体状态的满映射，也要求由于本体状态存在才使载体状态成为客观现实，凸显了信息兼具数学映射与物理实现双重属性的基本特征[24—26]。

2.2 状态的形式化表达

关于信息的公设强调本体的状态使能映射到载体的状态。因此，建立本体和载体状态的形式化表达尤为重要。

定义 1（高阶形式逻辑系统） 设形式系统 $\mathcal{L}$ 用以下符号库表示：

变元 $x_1, x_2, \dots$;

个体常元 $a_1, a_2, \dots$;

函数 $f_1^1, f_2^1, \dots, f_1^2, f_2^2, \dots, f_1^3, f_2^3, \dots$;

括号 $()$;

谓词 $A_1^1, A_2^1, \dots, A_1^2, A_2^2, \dots, A_1^3, A_2^3, \dots$;

联结词“非” $\sim$, 联结词“蕴含” $\rightarrow$;

量词“所有” $\forall$ 。

$\mathcal{L}$ 中的一个项定义如下：

(1) 变元和个体常元是项；

(2) 如果 $f_i^n (n > 0, i > 0)$ 是 $\mathcal{L}$ 中的函数，且 $t_1, \dots, t_n$ 都是 $\mathcal{L}$ 中不包含 f_i^n 自身的项，则 $f_i^n(t_1, \dots, t_n)$ 也是 $\mathcal{L}$ 中的项；

(3) 如果 $A_i^n (n > 0, i > 0)$ 是 $\mathcal{L}$ 中的谓词，且 $t_1, \dots, t_n$ 都是 $\mathcal{L}$ 中不包含 A_i^n 自身的项，则 $A_i^n(t_1, \dots, t_n)$ 也是 $\mathcal{L}$ 中的项；

(4) $\mathcal{L}$ 中所有项组成的集都按 (1)、(2) 和 (3) 生成。

由此可以定义 $\mathcal{L}$ 中的原子公式：

如果 $A_i^n (n > 0, i > 0)$ 是 $\mathcal{L}$ 中的一个谓词并且 $t_1, \dots, t_n$ 都是 $\mathcal{L}$ 中不包含 A_i^n 自身的项，则 $A_i^n(t_1, \dots, t_n)$ 是 $\mathcal{L}$ 中的一个原子公式。

进而可以定义 $\mathcal{L}$ 中的合式公式：

(5) $\mathcal{L}$ 中的每个原子公式是 $\mathcal{L}$ 中的合式公式；

(6) 如果 $\mathcal{A}$ 和 $\mathcal{B}$ 是 $\mathcal{L}$ 中的合式公式，则 $\sim \mathcal{A}$, $\mathcal{A} \rightarrow \mathcal{B}$, $(\forall x_i)\mathcal{A}$, $(\forall f_i^n)\mathcal{A}$, $(\forall A_i^n)\mathcal{A}$ 都是 $\mathcal{L}$ 中的合式公式，其中 x_i , f_i^n , A_i^n 分别是 $\mathcal{L}$ 中的变元、函数和谓词，且其中任何谓词或函数符号都不直接或间接作用于自身；

(7) $\mathcal{L}$ 中所有合式公式组成的集都按 (5) 和 (6) 生成。

这里定义的高阶形式逻辑系统 $\mathcal{L}$ ，基于谓词逻辑的灵活性可以描述各种状态间的关系与约束，通过递归结构与高阶量化能力支持复杂数据与行为的嵌套定义，并且强调任何谓词或函数都不直接或间接作用于自身以避免自指悖论。

出于简约性考虑， $\mathcal{L}$ 中没有定义联结词“合取” $\wedge$ 、“析取” $\vee$ 、“当且仅当” $\leftrightarrow$ 和量词“存在” $\exists$ 等常用的逻辑符号。但数理逻辑理论证明，这些符号完全等价于现有若干符号的组合应用，并且能够使很多公式更为简洁[27]。因此后续讨论中它们也作为被定义符号而引进。

定义 2（形式系统的解释） 形式系统 $\mathcal{L}$ 在论域 D_E 上的解释 E 是： D_E 为非空集，一个特定元素集 $(\widehat{a_1}, \widehat{a_2}, \dots)$ ，一个在 D_E 上的函数集 $(\widehat{f_i^n}, n > 0, i > 0)$ ，一个在 D_E 和特定元素集上的关系集 $(\widehat{A_i^n}, n > 0, i > 0)$ 。

按照这个定义， $\mathcal{L}$ 的变元 $x_1, x_2, \dots$ 将被解释为对象。集 D_E 将是假定变元在其上取值的对象领域。特定元素的集合由 $\mathcal{L}$ 中个体常元所代表的特殊对象组成。函数和关系将是对 $\mathcal{L}$ 中函数字母和谓词字母的具体解释。

定义 2 旨在尽可能以形式化方式描述主观和客观世界中对象所具有的状态，以支持严谨量化的数学处理和分析。这里的状态实际上是指对象的界定、性质、形态、取值、关系和演变规则以及它们之间的逻辑组合等。为此，建立以下公设：

公设 5（对象的状态表达） 对象集合在特定时间集合上的状态集合由其定义、性质、形态、取值、关系和演变规则以及它们之间的逻辑组合构成，且满足：

- (1) 定义不变性：任一对象在整个论域中的定义保持不变。
- (2) 性质可表达：任一对象在整个论域中的性质可通过谓词表达。
- (3) 形态可表达：任一对象在整个论域中的形态可通过函数或谓词表达。
- (4) 取值可表达：任一对象在整个论域中的取值可通过函数或谓词表达。
- (5) 关系可表达：任意多对象之间的关系在整个论域中可通过谓词表达。
- (6) 规则可表达：任意对象及其性质、形态、取值和关系在整个论域中的演变规则可通过时间参数化和蕴含联结词表达，且不包含隐含的自指结构。
- (7) 逻辑自洽性：状态集合中不存在逻辑矛盾的元素。
- (8) 逻辑封闭性：状态集合的逻辑闭包仍包含于状态集合。

定理 1（状态的形式化表达） 设对象集合 X 在特定时间集合 T 的状态集合 $S(X, T)$ 满足形式系统 $\mathcal{L}$ 的表达方式和公设 5，则其一定可以表示为 $\mathcal{L}$ 中某一合式公式集合在论域 $X \times T$ 上的一个解释，并且还是逻辑自洽和封闭的。

证明： 根据公设 5，我们逐项分析 $S(X, T)$ 的构成：

按照 (1)，对于任一对象 $x \in X$ ， $t \in T$ ，总有定义谓词 **Definition**，**Definition**(x, t, P) 恒为真，其中 P 为定义的具体内容。即

$$(\forall x \forall t ((\exists! P) \wedge \text{Definition}(x, t, P))) \in S(X, T) \quad (2.1)$$

这就表明 $S(X, T)$ 非空。

按照 (2) — (5)， X 在 T 的状态集合中的性质、形态、取值和关系等内容都可以通过 $S(X, T)$ 中的函数和谓词等得以表达，详细过程不再赘述。

按照 (6)，我们可以用 $A_1(x_1, t_1, f_1(x_1, t_1))$ 和 $A_2(x_2, t_2, f_2(x_2, t_2))$ 表达 X 在 T 的状态集合关于对象 $x_1, x_2 \in X$ ，时间 $t_1, t_2 \in T (t_1 < t_2)$ 和函数 $f_1(x_1, t_1), f_2(x_2, t_2)$ 的两个性质、形态、取值或关系，并且所有 $A_1(x_1, t_1, f_1(x_1, t_1))$ 符合从 t_1 到 t_2 演变为某一 $A_2(x_2, t_2, f_2(x_2, t_2))$ 的规则，即

$$(\forall A_1 \forall f_1 \forall x_1 A_1(x_1, t_1, f_1(x_1, t_1)) \rightarrow \exists A_2 \exists f_2 \exists x_2 A_2(x_2, t_2, f_2(x_2, t_2))) \in S(X, T) \quad (2.2)$$

即 $S(X, T)$ 能够通过递归方式表达状态演变规则。同样对于更复杂的情形，按照 (6) 也能表明 $S(X, T)$ 的可表达性。

因为 $\mathcal{L}$ 的定义已使其具有完整的合式公式逻辑组合能力，因此 $S(X, T)$ 能够表达状态集合元素的所有逻辑组合。

再由 (7)， $S(X, T)$ 是逻辑自洽的。

最后按照 (8)， $S(X, T)$ 的逻辑闭包仍包含于 $S(X, T)$ ，即所有从 $S(X, T)$ 中合式公式通过逻辑蕴含导出的公式仍然属于 $S(X, T)$ ，表明了 $S(X, T)$ 的逻辑封闭性。

综上，对象集合 X 在特定时间集合 T 的状态集合 $S(X, T)$ 非空，就是 $\mathcal{L}$ 中某一合式公式集合在论域 $X \times T$ 上的一个解释，并且还是逻辑自洽和封闭的。定理证毕。

定理 1 表明利用高阶逻辑系统 $\mathcal{L}$ 几乎可以描述世间所有事物的状态。[28]也从理论和实例分析表明 $\mathcal{L}$ 普遍适用于数学、经济学、计算机和信息系统、自然语言等学科的状态表达。这就为跨学科研究提供了统一的状态建模理论，对人工智能、知识工程和复杂系统等各个领域都有重要意义。

2.3 有限自动状态

利用状态的形式化表达，我们可以深入研究有限自动机这一特殊且重要的对象。

定理 2（有限自动机的状态表达） 对于任何有限自动机 M ，都能以其为对象，通过状态集合 $S(M, T)$ ，描述其在相关时间集合 T 上的输入、输出和状态转移行为。

证明： 设 $M = \langle Q, R, U, \delta, \lambda \rangle$ 为一有限自动机， $T = \{t_i | i = 1, \dots, n_T\}$ 为 M 发生输入、输出和状态转移行为且按序增加的时间集合。根据有限自动机理论， M 中 $Q = \{q_{t_i} | t_i = 1, \dots, n_T - 1\}$ ， $R = \{r_{t_i} | t_i = 2, \dots, n_r\}$ 和 $U = \{u_{t_i} | t_i = 1, \dots, n_u\}$ 都是非空有限集，分别为 M 的输入集、输出集和状态集； δ 是 $U \times Q$ 到 U 的映射，是 M 的下一状态函数； λ 是 $U \times Q$ 到 R 的单值映射，是 M 的输出函数[29]。

对于每个状态 $u_{t_i} \in U$ ，定义状态谓词 $State^3$ ，公式

$$\varphi_U(t_i) = State^3(M, t_i, u_{t_i}) \quad (2.3)$$

表示 M 在 t_i 时刻的状态为 u_{t_i} ；

对于每个输入 $q_{t_i} \in Q$ ，定义输入谓词 $Input^3$ ，公式

$$\varphi_I(t_i) = Input^3(M, t_i, q_{t_i}) \quad (2.4)$$

表示 M 在 t_i 时刻的输入为 q_{t_i} ；

对于每个输出 $r_{t_i} \in R$ ，定义输出谓词 $Output^3$ ，公式

$$\varphi_O(t_i) = Output^3(M, t_i, r_{t_i}) \quad (2.5)$$

表示 M 在 t_i 时刻的输出为 r_{t_i} ；

对于转移函数 δ ，其实现机制可用合式公式 $\varphi_\delta(t_i)$ 表示为：

$$\varphi_\delta(t_i) = \varphi_U(t_i) \wedge \varphi_I(t_i) \rightarrow \exists \delta(u_{t_i}, q_{t_i}) \wedge Eq^2(u_{t_{i+1}}, \delta(u_{t_i}, q_{t_i})) \wedge \varphi_U(t_{i+1}) \quad (2.6)$$

其中，二元谓词 $Eq^2(x, y)$ 表示 x 等于 y 。

对于输出函数 λ ，其实现机制可用合式公式 $\varphi_\lambda(t_i)$ 表示为：

$$\begin{aligned} \varphi_\lambda(t_i) = \varphi_U(t_i) \wedge \varphi_I(t_i) \rightarrow \\ \exists \lambda(u_{t_i}, q_{t_i}) \wedge Eq^2(r_{t_{i+1}}, \lambda(u_{t_i}, q_{t_i})) \wedge IsElementof^2(r_{t_{i+1}}, R) \wedge \varphi_O(t_{i+1}) \end{aligned} \quad (2.7)$$

其中，二元谓词 $IsElementof^2(x, X)$ 表示 x 是集合 X 的元素。

这样

$$S(M, T) = Cn(\{\varphi_U(t_i)\} \cup \{\varphi_I(t_i)\} \cup \{\varphi_O(t_i)\} \cup \{\varphi_\delta(t_i)\} \cup \{\varphi_\lambda(t_i)\}) \quad (2.8)$$

就能描述 M 在时间集合 T 上的输入、输出和状态转移行为。其中 $Cn(X)$ 表示合式公式集合 X 的逻辑闭包。定理证毕。

可见，有限自动机都能形式化表达为定理 1 给出的状态集合。并且这个状态还具有自动转换的特性。因此我们提出：

定义 3（有限自动状态） 若状态 $S(X, T)$ 表达一有限自动机，则称其为有限自动状态。

包含计算系统在内，几乎所有工作中的信息系统都是数学意义的有限自动机。因而有限自动状态的概念对于研究机器学习具有非常重要的意义。

2.4 信息的定义

[22]已经证明, 满足公设 1-4 的充分必要条件是分别存在非空的信息本体集合 o , 本体发生时间集合 T_h , 本体状态集合 $S_o(o, T_h)$ 以及客观载体集合 c , 载体反映时间集合 T_m , 载体反映集合 $S_c(c, T_m)$, 使 $S_o(o, T_h)$ 使能映射到 $S_c(c, T_m)$, 记此映射为 I , 简记为 $I = \langle o, T_h, S_o, c, T_m, S_c \rangle$, 称之为信息的六元组模型。

由此可以得到关于信息的一般定义:

定义 4 (信息的数学定义) 设 $\mathbb{O}$ 表示客观世界集合, $\mathbb{S}$ 表示主观世界集合, $\mathbb{T}$ 为时间集合, $\mathbb{O}$ 、 $\mathbb{S}$ 和 $\mathbb{T}$ 中的元素可以根据论域的特定要求进行合适的划分。本体 $o \in 2^{\mathbb{O} \cup \mathbb{S}}$ 、发生时间 $T_h \in 2^{\mathbb{T}}$ 、 o 在 T_h 上的状态集合 $S_o(o, T_h)$ 、载体 $c \in 2^{\mathbb{O}}$ 、反映时间 $T_m \in 2^{\mathbb{T}}$ 和 c 在 T_m 上的反映集合 $S_c(c, T_m)$ 均为非空集, 信息 I 便是 $S_o(o, T_h)$ 到 $S_c(c, T_m)$ 的使能映射, 即:

$$I: S_o(o, T_h) \rightrightarrows S_c(c, T_m) \quad (2.9)$$

或

$$I(S_o(o, T_h)) \supseteq S_c(c, T_m) \quad (2.10)$$

所有信息 I 的集合称为信息空间, 记为 $\mathfrak{I}$, 是客观世界的三大构成要素之一。

(2.9) 和 (2.10) 用符号 $\rightrightarrows$ 和 $\supseteq$ 表示使能映射, 与常用于映射的符号 $\rightarrow$ 和 $=$ 有所区别, 表示信息的本体状态不仅在数学上建立了与载体状态之间的满映射关系, 还在物理上形成了前者与后者之间的因果使能关系。这已经表明载体状态集合 $S_c(c, T_m)$ 皆由本体状态集合 $S_o(o, T_h)$ 引发, 因此使能映射 I 一定是满映射。同时为了尽可能简化问题, 后续研究中如未特别强调, 我们都假定 I 是单值映射。

可见, 对于任一 $\psi \in S_c(c, T_m)$, 都存在至少一个 $\varphi \in S_o(o, T_h)$, 使得 $I(\varphi) \supseteq \psi$ 。此时我们记为 $I^{-1}(\psi) \ni \varphi$ 或 $\varphi \in I^{-1}(\psi)$, 以保持使能方向的一致性。

现实情况中当然存在很多 I 为一一映射的情况, 这将使问题大为简明。

定义 5 (可还原信息) 设信息 $I = \langle o, T_h, S_o, c, T_m, S_c \rangle$ 还是 $S_o(o, T_h)$ 到 $S_c(c, T_m)$ 的一一映射, 即 I 可逆。因而存在逆映射 I^{-1} , 使得 $I^{-1}(S_c(c, T_m)) \subseteq S_o(o, T_h)$ 。这时我们称信息 I 可还原, 也称 $S_o(o, T_h)$ 是信息 I 的还原态。

特别地, 鉴于信息对于信息系统的重要意义, 我们提出:

定义 6 (信息系统中的信息) 设 $I = \langle o, T_h, S_o, c, T_m, S_c \rangle$ 为可还原信息, 若 $S_o(o, T_h)$ 为有限自动状态, 则称 c 是一个信息系统, I 是信息系统 c 中的信息。

对于可还原信息, 我们也用 $I^{-1}(S_c(c, T_m)) \subseteq S_o(o, T_h)$ 或 $I^{-1}: S_c(c, T_m) \hookrightarrow S_o(o, T_h)$ 表示使能方向的一致性。由于可还原信息中载体无损承载了本体的完整状态, 自然是信息研究中的理想情形。若 I 非一一映射, 我们总能把 $S_o(o, T_h)$ 中映像相同的元素划入同一等价类, 然后以等价类规约的方式建立各个等价类向 $S_c(c, T_m)$ 的使能映射, 它满足一一映射的条件, 因而可以被认为实现了信息 I 的可还原化。这就有:

命题 1 (信息的可还原规约) 对于信息 $I = \langle o, T_h, S_o, c, T_m, S_c \rangle$, 设 $S_o(o, T_h)$ 上的等价关系 R 为 $xRy \leftrightarrow I(x) = I(y)$ 。则能够建立唯一的 $S_o(o, T_h)/R$ 向 $S_c(c, T_m)$ 的使能映射 $\hat{I}: S_o(o, T_h)/R \rightrightarrows S_c(c, T_m)$, 并且 $\hat{I}$ 是一一映射, 即 $\hat{I}$ 可还原。此时称 $\hat{I}$ 是信息 I 的可还原规约。

当 I 就是可还原信息时， $S_o(o, T_h)/R = S_o(o, T_h)$ ， $\hat{I} = I$ 。可还原信息和信息的可还原规约将为简化问题提供非常重要的帮助。

[24]是公认的信息论奠基之作。该文提出“ $H = -\sum p_i \log p_i$ 在信息论中作为信息的度量起着核心作用”。其讨论的信息事实上也包括信源和信道两个部分。信源 Is 是信息的起源，它分别以 $p_1, p_2, \dots, p_n$ 的概率发生事件 $e_1, e_2, \dots, e_n$ 。信道 Ic 是传递信息的通道，当信源 Is 在时间 T_h 发生某一事件 e_i 后， Ic 就在时间区间 $(T_h, T_h + \tau)$ 接收并传送相应的编码 c_i ($1 \leq i \leq n, \tau > 0$)。我们定义三元谓词 $IsEvent^3(Is, T_h, e_i)$ 表示信源 Is 在时间 T_h 发生了事件 e_i ，定义三元谓词 $IcCode^3(Ic, T_m, c_i)$ 表示信道 Ic 在时间区间 T_m 接收并传送了编码 c_i 。则

$$S_{Is}(Is, T_h) = IsEvent^3(Is, T_h, e_i) \quad (2.11)$$

和

$$S_{Ic}(Ic, (T_h, T_h + \tau)) = IcCode^3(Ic, (T_h, T_h + \tau), c_i) \quad (2.12)$$

分别就是本体 Is 在发生时间 T_h 的状态和载体 Ic 在反映时间 $(T_h, T_h + \tau)$ 的状态 ($1 \leq i \leq n$)。而

$$I: S_{Is}(Is, T_h) \Rightarrow S_{Ic}(Ic, (T_h, T_h + \tau)) \quad (2.13)$$

就是由信源状态 $S_{Is}(Is, T_h)$ 向信道状态 $S_{Ic}(Ic, (T_h, T_h + \tau))$ 的使能映射。并且由(2.12)式中的 c_i 我们能够还原(2.11)式中的 e_i ，因而 I 还是可还原信息。可见，虽然 Shannon 信息论针对 $S_{Is}(Is, T_h)$ 的各种概率分布，研究 $S_{Ic}(Ic, (T_h, T_h + \tau))$ 的复杂实现机理，形成了内涵丰富、枝繁叶茂的理论体系。OIT 关于信息的定义能够涵盖 Shannon 最本原的信息模型，表明其具有更加广泛普适的应用范围。

2.5 有噪声信息

很多研究者怀疑可还原信息的意义。因为现实世界中的信息往往不可能完全还原。这是因为信息实现的过程中，很难保证本体状态不受噪声或其它外部因素的影响。由此引入以下概念：

定义 7 (有噪声信息) 对于信息 $I = \langle o, T_h, S_o, c, T_m, S_c \rangle$ ，设 $\widetilde{S}_o(o, T_h)$ 为本体 o 在发生时间 T_h 的另一状态， $\widetilde{S}_c(c, T_m)$ 为载体 c 在反映时间 T_m 的另一状态，

$$\widetilde{S}_o(o, T_h) = S_o(o, T_h) \setminus S_o^-(o, T_h) \cup S_o^+(o, T_h) \quad (2.14)$$

其中状态 $S_o^-(o, T_h)$ 称为 $S_o(o, T_h)$ 的损失状态，满足 $S_o^-(o, T_h) \subseteq S_o(o, T_h)$ 且未被使能映射到 $\widetilde{S}_c(c, T_m)$ ，状态 $S_o^+(o, T_h)$ 称为 $S_o(o, T_h)$ 的叠加状态，满足 $S_o(o, T_h) \cap S_o^+(o, T_h) = \emptyset$ 且被使能映射到 $\widetilde{S}_c(c, T_m)$ 。则称使能映射：

$$\tilde{I}: \widetilde{S}_o(o, T_h) \Rightarrow \widetilde{S}_c(c, T_m) \quad (2.15)$$

为信息 I 的有噪声信息，而 I 是其自身的无噪声信息。

可见有噪声信息 $\tilde{I}$ 也是一个信息。它与信息 I 具有相同的本体、发生时间、载体和反映时间。但两者的本体和载体状态虽可能有重叠，却不完全相同。进一步分析(2.14)式可知，信息 I 与 $\tilde{I}$ 完全对称，因而有：

命题 2 (有噪声信息对称性) 若信息 $\tilde{I}$ 是信息 I 的有噪声信息，则信息 I 也是信息 $\tilde{I}$ 的有噪声信息。

事实上，无论对于信息 I 还是其有噪声信息 $\tilde{I}$ ，单纯从使能映射的角度都不难验证其可还原性或可还原规约。但实际场景中，对于信息 I 往往很难获得它的无噪声状态，只能借助其某一有噪声信息 $\tilde{I}$ 开展研究。并且 $\tilde{I}$ 的还原态不是 I 的还原态，这是研究者怀疑信息可

还原性的根本原因。但在无法直接获得无噪声信息 I 的条件下，我们可以根据 $\tilde{I}$ 以及 $S_o(o, T_h)$ 、 $S_o^-(o, T_h)$ 和 $S_o^+(o, T_h)$ 三者之间的关系，逼近 I 的理想状态。这就是引入有噪声信息概念的实用意义。

Shannon 也是先由理想的无噪声信道入手，然后讨论更为复杂的有噪声信道情形[24]。此时 Shannon 假设信源产生的信号为 S ，信道噪声为 N ，而信道输出为

$$E = f(S, N) \quad (2.16)$$

其中 f 为信道对传送信号的失真函数。

按照 OIT 的观点，这里信号 S 就是信息 I 的本体状态 $S_{Is}(Is, T_h)$ ，对于信号 S 和噪声 N 的失真函数 $f(S, N)$ 等效于从信源状态 $S_{Is}(Is, T_h)$ 中移除某一状态 $S_{Is}^-(Is, T_h)$ 并添加另一状态 $S_{Is}^+(Is, T_h)$ ，而 E 就是信道的输出状态 $\widetilde{S}_{Ic}(Ic, (T_h, T_h + \tau))$ ， $(T_h, T_h + \tau)$ 是信道输出的时间区间。因而

$$\tilde{I}: \widetilde{S}_{Is}(Is, T_h) \Rightarrow \widetilde{S}_{Ic}(Ic, (T_h, T_h + \tau)) \quad (2.17)$$

就是

$$I: S_{Is}(Is, T_h) \Rightarrow S_{Ic}(Ic, (T_h, T_h + \tau)) \quad (2.18)$$

的有噪声信息，其中 $\widetilde{S}_{Is}(Is, T_h) = S_{Is}(Is, T_h) \setminus S_{Is}^-(Is, T_h) \cup S_{Is}^+(Is, T_h)$ 。

这样我们也用有噪声信息的定义阐释了[24]关于有噪声信道的信息模型，表明 OIT 完全适用于通常有噪声条件下信息研究的复杂情形。

3 机器学习理论元框架（MLT-MF）

机器学习的核心是模型，它是典型的有限自动机，也是一个信息系统。因而每一次学习以及其后的测试或应用过程都依赖于其中的信息，一般可以归纳分解为四个主要的信息实现和作用步骤。

第一步：训练信息 It 的实现，就是将 It 的本体状态使能映射到模型接口这个载体上；

第二步：模型中具有学习能力的原有信息 Io 与训练信息 It 融合，优化为兼具问题回答能力和学习能力的更新信息 Iu ；

第三步：用户输入信息 Iq 的实现，也是将 Iq 的本体状态使能映射到模型接口这个载体上；

第四步：模型中的更新信息 Iu 处理输入信息 Iq ，并根据用户需求输出信息 Ir ，同时 Iu 又经过学习优化为 Iv 。

3.1 训练信息的形式化建模与实现

根据通用的信息形式化定义，训练信息实现的形式化表达非常简单。即：

$$It = \langle ot, Tt_h, S_{ot}, ct, Tt_m, S_{ct} \rangle \quad (3.1)$$

$$S_{ot}(ot, Tt_h) = \{\varphi t | \varphi t \text{ 为关于 } ot \text{ 和 } Tt_h \text{ 的合式公式}\} \quad (3.2)$$

$$S_{ct}(ct, Tt_m) = \{\psi t | \psi t \text{ 为关于 } ct \text{ 和 } Tt_m \text{ 的合式公式}\} \quad (3.3)$$

其中，训练信息 It 的本体集合 ot 是 It 所蕴含某类知识、命题、事实或样本等状态直接从属的主体对象，发生时间集合 Tt_h 是本体 ot 发生相应状态的时间阶段，本体状态集合 $S_{ot}(ot, Tt_h)$ 是本体 ot 在时间 Tt_h 上发生的状态，是针对本体 ot 、面向业务和事件的语义断言、性质或事实。例如，假设训练信息 It 表达“用户 X 在 T_1 时段内活跃，并且设备 Y 在 T_2 时刻的温度大于 60°C ”这个事实，我们可以得到本体 $ot = \{X, Y\}$ ，发生时间 $Tt_h = \{T_1, T_2\}$ ，

谓词 $Active^2(x, t)$ 表示“ x 在 t 时段很活跃”，函数 $f_{temp}^2(x, t)$ 表示“ x 在 t 时间的温度”，谓词 $Gt^2(x, y)$ 表示“ x 大于 y ”，则有本体状态 $S_{ot}(ot, Tt_h) = \{\varphi t\} = \{Active^2(X, T_1) \wedge Gt^2(f_{temp}^2(Y, T_2), 60C^\circ)\}$ 这样的形式化表达。可见， $S_{ot}(ot, Tt_h)$ 并不是简单的数据罗列，而是对知识、事实、特征等内容的形式化表达和归纳，同时它也还未实现信息能被接收、感知和处理的要求。

为此，就需要类似于硬件接口和 API 软件等的模型训练数据接口 ct 作为训练信息 It 的载体，以工程系统可接收的方式承接 $S_{ot}(ot, Tt_h)$ 的内容。反映时间集合 Tt_m 是载体反映信息 It 的时间阶段，载体反映集合 $S_{ct}(ct, Tt_m)$ 是载体 ct 在时间 Tt_m 上反映信息 It 的一种具体、完整、可操作的实现状态集合，其元素是一系列被 φt 使能反映、相互之间具有对应逻辑和语义关系的形式化合式公式 ψt 。还是以上段列举的训练信息 It 为例，当 $S_{ot}(ot, Tt_h)$ 在 T_3 时段被使能映射到 ct 时，就有 $S_{ct}(ct, T_3) = \{\psi t\} = \{(TrainTable[user_{id} = X][time = T_1] = Active) \wedge (TrainTable[device_{id} = Y][time = T_2] = Tgt60)\}$ ，这里Active和Tgt60在训练数据表中分别被定义为“用户活跃”和“设备温度大于 60 C°”两种特定状态。可见，合式公式 ψt 的形式虽然简单，已经全然反映了 φt 的内容涵义。实际情形中 ψt 可为表格、向量、JSON、API 参数等多种工程数据接口或硬件设备状态的形式化表达，可能具有更加复杂的形式，还可以进一步包括数据约束、格式校验等内容，确保 $S_{ct}(ct, T_3)$ 的合法性与鲁棒性。

还必须强调的是， $S_{ot}(ot, Tt_h)$ 只有经过 It 的使能映射才实现 $S_{ct}(ct, T_3)$ 。 It 既体现了前者到后者的结构性语义映射，也表明前者为因，后者是果，并且由前者到后者需要耗费人力、时间和能量等资源才能得以实现。但也正是依赖于这样的使能映射，才使本体蕴含的知识、事实、规则等形式化表达以结构化、标准化、可机读的方式进入模型，成为后续模型学习可以利用的信息。

实际工程中，训练信息的实现过程往往由于离散化、属性剪裁和噪声叠加等原因需要考虑有噪声信息 $\tilde{It}$ 。但为便于理论分析，目前我们还是仅针对理想的无噪声情形，复杂的有噪声条件留待后续深入研究。

3.2 模型学习的形式化建模与实现

基于信息流转应用的视角，人工智能模型 M 的学习就是其中的原有信息 Io 通过学习训练信息 It ，优化为更新信息 Iu 的过程。前段我们已经描述了训练信息 It 的模型与实现。现在需要关注 M 中 Io 和 Iu 的模型与实现。为此参照式 (3.1) — (3.3) 对于 It 的定义，只是将其中字符 t 对应的位置分别替换为 o 和 u ，就能得到 Io 和 Iu 的定义。需要说明的是，由于 Io 和 Iu 分别代表 M 学习前后所具有的信息，因此它们的载体 co 和 cu 可视为完全相同，即可直接以 M 指代，更严格地可以说就是 M 运行的计算系统。即

$$co = cu = M \quad (3.4)$$

在 M 学习的环节，最重要的就是其中的原有信息 Io ，它除与 It 一样蕴含和表达一些知识、命题、事实或样本等之外，不同之处就是它自身具有学习规则，即通过学习优化为更新信息的规则。

定义 8（具有学习能力的状态）在定义 1 所述形式系统中补充特定的函数和谓词定义：

函数： $f_{learn}^2(X, Y)$ 表示集合 X 向 Y 学习后形成的合式公式；

谓词：IsSubsetOf²(X, Y) 表示集合X是集合Y的子集，Learnable²(X, Y) 表示状态集合X能够向状态集合Y学习，IsFormulaOf²(x, X) 表示x是状态集合X中的合式公式，Eq²(x, y)表示x等于y。

则对于对象x在时间T_x上的状态Sx(x, T_x)，对象y在时间T_y上的状态Sy(y, T_y)，对象z在时间T_z上的状态Sz(z, T_z)，当合式公式φx ∈ Sx(x, T_x)，sup(T_x, T_y) ≤ inf(T_z)且

$$\begin{aligned} \varphi x = (\forall \Phi x)(\forall \Phi y)(\text{IsSubsetOf}^2(\Phi x, Sx(x, T_x)) \wedge \text{IsSubsetOf}^2(\Phi y, Sy(y, T_y)) \wedge \\ \text{Learnable}^2(\Phi x, \Phi y)) \rightarrow \exists \varphi z(\text{IsFormulaOf}^2(\varphi z, Sz(z, T_z)) \wedge \text{Eq}^2(\varphi z, f_{learn}^2(\Phi x, \Phi y))) \end{aligned} \quad (3.5)$$

时，我们称Sx(x, T_x)具有向Sy(y, T_y)学习并优化到Sz(z, T_z)的能力。

φx实际上是状态Sx(x, T_x)的学习规则，表明只要Sy(y, T_y)满足Sx(x, T_x)的学习条件，就能在Sz(z, T_z)中新增合式公式集合{φu | φu = f_{learn}²(Φx, Φy), Φx ⊆ Sx(x, T_x), Φy ⊆ Sy(y, T_y)}，使Sz(z, T_z)成为Sx(x, T_x)的优化。

另一方面，学习能力的继承也是模型学习的重要特性。实际上，为继承Sx(x, T_x)的学习能力，只需在Sz(z, T_z)中保留学习规则φx，仅对其部分符号进行适应性替代，即φz ∈ Sz(z, T_z)，且

$$\begin{aligned} \varphi z = (\forall \Phi z)(\forall \Phi y)(\text{IsSubsetOf}^2(\Phi z, Sz(z, T_z)) \wedge \text{IsSubsetOf}^2(\Phi y, Sy(y, T_y)) \wedge \\ \text{Learnable}^2(\Phi z, \Phi y)) \rightarrow \exists \varphi w(\text{IsFormulaOf}^2(\varphi w, Sw(w, T_w)) \wedge \text{Eq}^2(\varphi w, f_{learn}^2(\Phi u, \Phi y))) \end{aligned} \quad (3.6)$$

就使Sz(z, T_z)具有向Sy(y, T_y)学习并优化到Sw(w, T_w)的能力。这就证明了：

命题 3（状态学习能力的继承） 若状态集合Sx(x, T_x)具有向Sy(y, T_y)学习并优化到Sz(z, T_z)的能力，则总能使Sz(z, T_z)继承Sx(x, T_x)的学习能力。

定义 9（模型学习的实现） 对于人工智能模型M的原有信息Io、训练信息It和更新信息Iu，若三者皆可还原，且本体状态S_{oo}(oo, To_h)具有向S_{ot}(ot, Tt_h)学习并优化到S_{ou}(ou, Tu_h)的能力，载体状态S_{co}(co, To_m)也具有向S_{ct}(ct, Tt_m)学习并优化到S_{cu}(cu, Tu_m)的能力，则称M能够实现信息Io向It学习并优化为Iu的能力，或简称M具有信息学习能力。

此时根据式（3.4），我们也可以记为S_{co}(co, To_m) = S_M(M, To_m)，S_{cu}(cu, Tu_m) = S_M(M, Tu_m)，表明学习过程只是模型M的状态优化过程。

至此，我们在理论上已经将模型学习能力完全内化为其中的状态属性，并且这些规则与模型的算法、结构和规模无关，为研究分析各类模型提供了统一普适的形式化基础框架。

作为一个实例，我们设Mn为一神经网络模型。按照常识，Mn中原有信息Io = {oo, To_h, S_{oo}, Mn, To_m, S_{co}}，oo是设计者心目中的模型，To_h是设计者设计Mn时的时间集合，

$$S_{oo}(oo, To_h) = \Phi o = \{Params(\theta) \wedge Architecture(L_1, \dots, L_n) \wedge Activation(A)\} \quad (3.7)$$

其中： $Params(\theta)$ 为参数集合 $\theta = \{W_1, b_1, \dots, W_n, b_n\}$ 的谓词表述， $Architecture(L_1, \dots, L_n)$ 为网络层结构的谓词表述， $Activation(A)$ 为激活函数集合 $A = \{a_1, \dots, a_n\}$ 的谓词表述。

信息 Io 的载体就是神经网络模型 Mn ， Io 的反映时间 To_m 就是 Mn 完成开发以后的时间， $S_{co}(co, To_m)$ 是公式(3.7)在 Mn 中的物理实现。

Mn 的训练信息 $It = \langle ot, Tt_h, S_{ot}, ct, Tt_m, S_{ct} \rangle$ ， ot 是实际世界中的某一对象集合， Tt_h 是对象集合发生状态 $S_{ot}(ot, Tt_h)$ 的时间集合，

$$S_{ot}(ot, Tt_h) = \Phi t = \{Batch(X, y) \wedge DimMatch(X, L_1) \wedge Differentiable(L)\} \quad (3.8)$$

其中： $Batch(X, y)$ 是输入数据 $X \in \mathbb{R}^{b \times d}$ 和标签 $y \in \mathbb{R}^{b \times k}$ 的谓词表达， $DimMatch(X, L_1)$ 是输入数据 X 与网络输入层 L_1 维度匹配的谓词表达， $Differentiable(L)$ 是损失函数 L 对参数 θ 可导的谓词表达。

可学习条件Learnable²谓词的定义是：

$$\begin{aligned} & \text{Learnable}^2(\Phi o, \Phi t) \\ &= \text{Dimensionality}(\Phi o, \Phi t) \wedge \text{ActivationFeasible}(\Phi o, \Phi t) \wedge \text{GradientExists}(\Phi t) \quad (3.9) \end{aligned}$$

其中： $\text{Dimensionality}(\Phi o, \Phi t) = DIM(X, b \times d) \wedge DIM(L_1, d)$ 是输入数据维度与网络输入层一致的谓词表达， $\text{ActivationFeasible}(\Phi o, \Phi t)$ 是对 $\forall a_i \in A, \text{Range}(X) \subseteq \text{Domain}(a_i)$ ，即输入数据在激活函数定义域内的谓词表达， $\text{GradientExists}(\Phi t)$ 是 $\exists \nabla_{\theta} L(\theta, X, y)$ ，即损失函数 L 对参数 θ 可导的谓词表达。

学习函数 f_{learn}^2 的定义是：

$$f_{learn}^2(\Phi o, \Phi t) = \{Params(\theta_{new}) \wedge Architecture(L_1, \dots, L_n) \wedge Activation(A)\} \quad (3.10)$$

其中 $\theta_{new} = \theta - \eta(\nabla_{\theta} L(\theta, X, y) + \beta v_{prev})$ 是神经网络模型的迭代公式。

由此，神经网络模型 Mn 的更新信息 $Iu = \langle ou, Tu_h, S_{ou}, Mn, Tu_m, S_{cu} \rangle$ 中，本体 ou 是设计者的思想与训练信息 It 的本体 ot 之并，发生时间 Tu_h 是 Mn 的训练时间集合，

$$S_{ou}(ou, Tu_h) = f_{learn}^2(\Phi o, \Phi t) \quad (3.11)$$

成为学习优化后的本体状态， Iu 的载体就是 Mn 自身，反映时间 Tu_m 是学习训练后的时间， $S_{cu}(Mn, Tu_m)$ 是 $S_{ou}(ou, Tu_h)$ 在 Mn 中的物理实现，理想情形当然满足使能映射的数学和物理要求，使得：

$$S_{cu}(Mn, Tu_m) = f_{learn}^2(\Psi o, \Psi t) \quad (3.12)$$

这里 Ψo 和 Ψt 分别是 Io 和 It 的两个载体状态集合内部合式公式的集合，是 Φo 和 Φt 在模型 Mn 及其接口上的物理实现。式(3.11)和(3.12)分别表明 Mn 的原有信息 Io 、训练信息 It 和更新信息 Iu 中，本体状态 $S_{oo}(oo, To_h)$ 具有向 $S_{ot}(ot, Tt_h)$ 学习并优化到 $S_{ou}(ou, Tu_h)$ 的能力，载体状态 $S_{co}(co, To_m)$ 也具有向 $S_{ct}(ct, Tt_m)$ 学习并优化到 $S_{cu}(cu, Tu_m)$ 的能力，故 Mn 能够实现 Io 向 It 学习并优化为 Iu 的能力。

需要说明的是，上述论述中部分内容没有完全采用严格的集合定义和合式公式表达，这不意味着它们不能转换为完全的集合定义和合式公式，而仅仅因为这样表示更加简明直观，同时也不会导致歧义。我们还能非常简单地按照上述程式，证明线性回归、决策树等更多典型的模型都具有信息学习能力，检验模型学习形式化建模的普适性。限于篇幅，本文不再赘述。

3.3 模型的信息处理与学习优化

模型学习的根本目的是为了更加正确地处理用户输入的信息 Iq ，反馈用户需要的高质量输出信息 Ir 。因为输入信息 Iq 的建模和实现与 It 几乎完全相同，3.1 节已做详细讨论，此处不再赘述。所以现在需要关注的是模型 M 中更新信息 Iu 处理 Iq ，输出 Ir ，同时又经过学习优化为 Iv 的过程。为此，仍参照式 (3.1) — (3.3) 的方式定义 Iq 、 Ir 和 Iv 。

定义 10（状态处理的形式化表达） 在定义 1 所述形式系统中补充特定的函数和谓词定义：

函数： $f_{process}^2(X, Y)$ 表示状态集合 X 处理状态集合 Y 后输出的合式公式集合；

谓词： $IsSubsetOf^2(X, Y)$ 表示集合 X 是集合 Y 的子集， $Processable^2(X, Y)$ 表示状态集合 X 能够处理状态集合 Y 并生成输出， $IsFormulaOf^2(x, X)$ 表示 x 是状态集合 X 中的合式公式， $Eq^2(x, y)$ 表示 x 等于 y 。

对于对象 z 在时间 T_z 上的状态 $Sz(z, T_z)$ 、对象 u 在时间 T_u 上的状态 $Su(u, T_u)$ 、对象 v 在时间 T_v 上的状态 $Sv(v, T_v)$ ，当合式公式 $\varphi z \in Sz(z, T_z)$ ， $\sup(T_z, T_u) \leq \inf(T_v)$ ，且

$$\varphi z = (\forall \Phi z)(\forall \Phi u)(IsSubsetOf^2(\Phi z, Sz(z, T_z)) \wedge IsSubsetOf^2(\Phi u, Su(u, T_u)) \wedge Processable^2(\Phi z, \Phi u)) \rightarrow \exists \varphi v(IsFormulaOf^2(\varphi v, Sv(v, T_v)) \wedge Eq^2(\varphi v, f_{process}^2(\Phi z, \Phi u))) \quad (3.13)$$

时，我们称 $Sz(z, T_z)$ 具有处理 $Su(u, T_u)$ 并输出 $Sv(v, T_v)$ 的能力。

可见，结合应用定义 8 和定义 10，我们也能够得到状态 $S_{ou}(ou, Tu_h)$ 处理并学习 $S_{oq}(oq, Tq_h)$ ，输出 $S_{or}(or, Tr_h)$ 又优化为 $S_{ov}(ov, Tv_h)$ 的形式化表达。由此可得：

定义 11（模型信息处理与学习优化的实现） 对于人工智能模型 M 的更新信息 Iu 、输入信息 Iq 、输出信息 Ir 和优化信息 Iv ，若四者皆可还原，且本体状态 $S_{ou}(ou, Tu_h)$ 具有处理并学习 $S_{oq}(oq, Tq_h)$ ，输出 $S_{or}(or, Tr_h)$ 又优化为 $S_{ov}(ov, Tv_h)$ 的能力，载体状态 $S_M(M, Tu_h)$ 也具有处理并学习 $S_{cq}(cq, Tq_h)$ ，输出 $S_{cr}(cr, Tr_h)$ 又优化为 $S_M(M, Tv_h)$ 的能力，则称 M 能够实现信息 Iu 处理并学习 Iq ，输出 Ir 并优化为 Iv 的能力，或简称 M 具有信息处理和学习优化能力。

对于上一节列举的神经网络模型 Mn ，可处理谓词 $Processable^2$ 为：

$$Processable^2(\Phi u, \Phi q) = InputMatch(\Phi u, \Phi q) \wedge ActivationFeasible(\Phi u, \Phi q) \wedge OutputDefined(\Phi u) \quad (3.14)$$

其中： $InputMatch(\Phi u, \Phi q) = DIM(\Phi q, b \times d) \wedge DIM(L_1, d)$ （ L_1 为 Φu 的输入层， b 为批次大小， d 为输入特征维度），是输入数据 Φq 维度与模型 Φu 输入层匹配的谓词表达； $ActivationFeasible(\Phi u, \Phi q) = (\forall a_i \in A) \wedge (Range(\Phi q) \subseteq Domain(a_i))$ （ A 为 Φu 的激活函数集合），是输入数据 Φq 的值域在模型 Φu 所有激活函数定义域内的谓词表达； $OutputDefined(\Phi u) = (\exists L_n \in Architecture(\Phi u)) \wedge DIM(L_n, k)$ （ L_n 为输出层， k 为输出维度）模型 Φu 的输出层已定义且维度合法的谓词表达。

处理函数 $f_{process}^2$ 为：

$$f_{process}^2(\Phi u, \Phi q) = f_{\Phi u}(\Phi q) \quad (3.15)$$

其中 $f_{\Phi u}$ 是依赖于 Mn 网络结构和参数能够显式表达的预测函数。

这样，神经网络模型 Mn 更新信息 Iu 的本体状态 $S_{ou}(ou, Tu_h) = \Phi u$ ，输入信息 Iq 的本体状态 $S_{oq}(oq, Tq_h) = \Phi q$ ，它们满足可处理条件 $Processable^2(\Phi u, \Phi q)$ ，因而其输出信息 Ir 的本体状态

$$S_{or}(or, Tr_h) = \Phi r = f_{process}^2(\Phi u, \Phi q) = f_{\Phi u}(\Phi q) \quad (3.16)$$

表明本体状态 $S_{ou}(ou, Tu_h)$ 具有处理 $S_{oq}(oq, Tq_h)$ 并输出 $S_{or}(or, Tr_h)$ 的能力。同样由于 Iu 、 Iq 和 Ir 皆为可还原信息，但三个本体状态都被使能映射到 Mn 及其接口时，三个载体状态中 $S_{Mn}(Mn, Tu_h)$ 也具有处理 $S_{cq}(cq, Tq_h)$ 并输出 $S_{cr}(cr, Tr_h)$ 的能力。可见 Mn 能够实现信息 Iu 处理 Iq 并输出 Ir 的能力。再根据 3.2， Mn 同时还能实现信息 Iu 学习 Iq 并优化为 Iv 的能力。

当然，我们也能用定义 8 和定义 10 的形式化架构很简单地证明线性回归、决策树等更多典型模型都兼具信息处理和学习优化能力。可见，这个架构通过公理化和形式化方法，超越了具体算法或模型，为机器学习理论提供了更加普适、统一、可解释的基础性元框架。

4 若干重要定理及证明

MLT-MF 适用于人工智能领域的一系列关键问题，能够提供很多独特而实用的重要理论规律和方法指南。

4.1 模型可解释性定理

像深度学习这样复杂模型的可解释性一直是人工智能领域的关键问题之一。但长期以来学界对模型可解释这一概念并未形成严格的数学表达[6, 30-32]。事实上，宇宙中任何现象的可解释性都应该是相对的。例如，开普勒三大定律解释了行星围绕太阳的运行规律，但对于不掌握初等数学知识的人而言，这样的解释并无意义。即使对于同一位学者，所能理解和接受的解释也会随着认识的拓展而变化。因此可解释性应该是依赖于对象和时间的一个概念。由此基于 MLT-MF 建立以下概念：

定义 12（模型可解释性） 设模型 M 在时间 Tm 上的状态为 $S_M(M, Tm)$ 。若存在对象集合 x 和时间集合 t ，形式系统 $\mathcal{L}$ 在论域 $x \times t$ 上的一个解释为 E ，使得 $S_M(M, Tm)$ 是 E 的一一使能映射，则称 M 在 $\mathcal{L}$ 中关于 x 和 t 可解释。

这样，我们利用形式系统 $\mathcal{L}$ 及其解释 E 定义了模型的可解释性。无论从人们对于“解释”一词的常识性认知，还是基于一系列命题刻画模型内在逻辑的理论需求，这样的定义都适合对人工智能可解释性的形式化解析和深入研究。

事实上，按照定理 1，一定存在模型 M 遵循的一个形式系统 $\mathcal{L}_M$ 和其中的合式公式集合在论域 $M \times Tm$ 上的解释 E ， $S_M(M, Tm) = E$ ，并且 $I: S_M(M, Tm) \rightrightarrows S_M(M, Tm)$ 是 $S_M(M, Tm)$ 到 $S_M(M, Tm)$ 的自映射，当然符合一一使能映射的条件。由此可得：

推论 1（模型的内在可解释性） 若模型 M 在时间 Tm 上的状态集合为 $S_M(M, Tm)$ ，则存在形式系统 $\mathcal{L}_M$ ， M 在 $\mathcal{L}_M$ 中关于 M 和 Tm 可解释。

正如黑格尔的著名论断“存在即合理”一样，模型的内在可解释性表明，它既然能够按照既定规则和程序处理输入信息并生成输出信息，自然具有其内在可解释的运行逻辑。但是这些逻辑可能高度隐含而抽象，因而很难被人们甚至是领域专家们所识别和理解。这样就突显出需要将其内在逻辑映射转化为更加直观和通用的逻辑系统之中，以便

于深入分解和剖析其中的因果机理，获得更加广泛的理解和认可，这才是模型可解释性的关键所在，也正是定义 12 提出在更加显性的形式系统之中解释模型内在逻辑的基本思想。

定理 3（模型可解释与信息可还原性） 设模型 M 中的信息为 $I = \langle o, T_h, S_o, M, T_m, S_M \rangle$ ，若信息 I 可还原，则存在形式系统 $\mathcal{L}$ ， M 在 $\mathcal{L}$ 中关于 o 和 T_h 可解释。

证明： 根据定义 5，因为 $I = \langle o, T_h, S_o, M, T_m, S_M \rangle$ 可还原，所以是一一使能映射，且 $I(S_o(o, T_h)) \supseteq S_M(M, T_m)$ 。

根据定理 3， $S_o(o, T_h)$ 是某一形式系统 $\mathcal{L}$ 中的合式公式集合在论域 $o \times T_h$ 上的一个解释，所以 $S_M(M, T_m)$ 就是这个解释的一一使能映射。

根据定义 12， M 在 $\mathcal{L}$ 中关于 o 和 T_h 可解释。定理证毕。

定理 3 充分体现了中国谚语“解铃还须系铃人”的涵义。因为直观上模型 M 的内在逻辑 $S_M(M, T_m)$ 源自于使其成为现实的本体状态 $S_o(o, T_h)$ ，因此用 $S_o(o, T_h)$ 解释 $S_M(M, T_m)$ 顺理成章。并且不同于 $S_M(M, T_m)$ 隐藏于模型 M 的内部， $S_o(o, T_h)$ 是以显式形式化表达的一系列合式公式，更加容易为人们，特别是专家所解读和理解。

4.2 伦理安全保障定理

伦理安全一直是人工智能领域备受关注的热点问题之一[33-36]，并且迄今为止也没有形成普遍认可且具有实践指导意义的严格定义[37-39]。基于 MLT-MF，我们能够通过伦理约束谓词嵌入，从理论上保障人工智能的伦理安全。为此首先需要针对公平性、隐私保护等基本伦理规范，确定相应的伦理约束谓词 $Ec(x)$ ，表示合式公式 x 符合伦理 Ec 的约束要求，由此建立伦理安全的形式化表达。

定义 13（伦理安全） 设状态 $S(x, t)$ 是形式系统 $\mathcal{L}$ 的一个合式公式集合关于论域 $x \times t$ 的一个解释， $Ec(y)$ 是 $\mathcal{L}$ 中的一个伦理约束谓词。若 $\forall \varphi \in S(x, t)$ ，都没有 $\varphi \rightarrow \sim Ec(\varphi)$ ，则称 $S(x, t)$ 关于伦理 Ec 安全。又设模型 M 能够实现输出信息 Ir 的能力，若 Ir 的本体和载体状态都关于伦理 Ec 安全，则称模型 M 关于伦理 Ec 安全。

定理 4（伦理安全保障） 设模型 M 能够实现信息 Iu 处理 Iq ，并输出 Ir 的能力。对于伦理约束谓词 $Ec(x)$ ，总能通过在 Iu 中嵌入 Ec 保障机制，实现 M 关于伦理 Ec 安全。

证明： 因为模型 M 能够实现信息 Iu 处理 Iq ，并输出 Ir 的能力，所以对于 Iu 、 Iq 和 Ir 的本体状态 $S_{ou}(ou, Tu_h)$ 、 $S_{oq}(oq, Tq_h)$ 和 $S_{or}(or, Tr_h)$ ，总有

$$\begin{aligned} \varphi u = (\forall \Phi u)(\forall \Phi q)(& \text{IsSubsetOf}^2(\Phi u, Sou(ou, Tu_h)) \wedge \text{IsSubsetOf}^2(\Phi q, Soq(oq, Tq_h)) \wedge \\ & \text{Processable}^2(\Phi u, \Phi q)) \rightarrow \exists \varphi r(\text{IsFormulaOf}^2(\varphi r, Sor(or, Tr_h)) \wedge \\ & \text{Eq}^2(\varphi r, f_{process}^2(\Phi u, \Phi q))) \quad (4.1) \end{aligned}$$

此时在 (4.1) 式中嵌入 $\sim (\varphi r \rightarrow \sim Ec(\varphi r))$ 条件，即

$$\begin{aligned} \varphi u = (\forall \Phi u)(\forall \Phi q)(& \text{IsSubsetOf}^2(\Phi u, Sou(ou, Tu_h)) \wedge \text{IsSubsetOf}^2(\Phi q, Soq(oq, Tq_h)) \wedge \\ & \text{Processable}^2(\Phi u, \Phi q)) \rightarrow \exists \varphi r(\text{IsFormulaOf}^2(\varphi r, S_{or}^{Ec}(or, Tr_h)) \wedge \\ & \text{Eq}^2(\varphi r, f_{process}^2(\Phi u, \Phi q)) \wedge (\sim (\varphi r \rightarrow \sim Ec(\varphi r)))) \quad (4.2) \end{aligned}$$

从而 $\forall \varphi r \in S_{or}^{Ec}(or, Tr_h)$ ，都有 $\sim (\varphi r \rightarrow \sim Ec(\varphi r))$ 。

这里 $S_{or}^{Ec}(or, Tr_h)$ 是通过 $S_{ou}(ou, Tu_h)$ 增加伦理 Ec 安全保障机制，去除了 $S_{or}(or, Tr_h)$ 中有违 Ec 的相关合式公式后形成的集合。根据定义 13，以其作为输出状态，即能实现关于伦理 Ec 安全。

因为 Iu 、 Iq 和 Ir 皆为可还原信息，它们的载体状态都是各自的本体状态在模型 M 上的使能映射，也就是物理实现。因而以 $S_{or}^{Ec}(or, Tr_h)$ 为本体状态的 Ir ，其载体状态 $S_M(M, Tr_m)$ 也必定关于伦理 Ec 安全，满足定义 13 模型 M 关于伦理 Ec 安全的条件。定理证毕。

定理 4 已经通过伦理约束谓词的嵌入，从理论上解决了所有模型的伦理安全问题。因而就没有再专门述及如何保证学习过程也能继承模型的伦理安全性。事实上，学习过程中模型伦理安全性的继承也可以通过模型算法附加专门的伦理约束正则项实现。这涉及更加具体的模型结构和算法设计，本文不再赘述。

总之，定义 13 和定理 4 首次构建了伦理安全的形式化理论框架，有别于基于启发式规则或统计指标的方法，填补了传统研究在数学基础与可验证性上的空白，能为伦理约束的自动化嵌入与验证提供具体的方法论指导。

4.3 模型泛化误差估计定理

提高模型泛化能力是机器学习最为重要的目标之一。但通常泛化误差难以事先进行理论分析和推导[40-44]，只能在模型完成训练后通过测试估计，并且很难依据有限样本的测试获得准确的误差指标。应用 MLT-MF，能够针对模型泛化误差的上界作出估计。

定理 5（模型泛化误差上界） 设模型 M 经过学习优化后的信息为 $Iu = \langle ou, Tu_h, S_{ou}, M, Tu_m, S_M \rangle$ ，输入信息 $Iq = \langle oq, Tq_h, S_{oq}, cq, Tq_m, S_{cq} \rangle$ ， $|S_{oq}|$ 和 $|S_{ou} \cap S_{oq}|$ 分别表示集合 S_{oq} 和 $S_{ou} \cap S_{oq}$ 的势， $p_q(x)$ 为 S_{oq} 服从的概率分布， $p_q(x)$ 在 $S_{ou} \cap S_{oq}$ 的概率总和为 P_{Mq} ，则当 S_{ou} 和 S_{oq} 服从定义于同一可测空间的概率分布， M 针对 S_{oq} 训练充分且对未知状态采用保守的均匀分布预测，处理 Iq 的泛化误差上界由 S_{oq} 及其与 S_{ou} 的状态重叠情况以及 $p_q(x)$ 在未重叠区域的熵决定，具体为：

$$\mathcal{E}_{\text{gen}} \begin{cases} \rightarrow 0, & \text{当 } S_{oq} \subset S_{ou} \\ \leq \sqrt{\frac{1}{2}((1 - P_{Mq}) \log((|S_{oq}| - |S_{ou} \cap S_{oq}|) / (1 - P_{Mq})) - H(p_q))}, & \text{当 } S_{oq} \neq S_{ou} \cap S_{oq} \end{cases} \quad (4.3)$$

其中 $\mathcal{E}_{\text{gen}}$ 为泛化误差上界， $H(p_q) = -\sum_{x \in S_{oq} \setminus S_{ou}} p_q(x) \log p_q(x)$ 为未重叠区域数据的熵。

证明： 设 $p_M(x)$ 是由模型 M 的结构和参数集合决定的概率分布， $Iu = \langle ou, Tu_h, S_{ou}, M, Tu_m, S_M \rangle$ 是 M 承载信息的集合，所以当 $x \in S_{ou}$ 时 $p_M(x) > 0$ ；当 $x \in S_{oq}$ 时 $p_q(x) > 0$ 。

因为 M 针对 S_{oq} 训练充分且对未知状态采用保守的均匀分布预测，因此 $p_M(x)$ 在重叠区域 $S_{ou} \cap S_{oq}$ 近似真实分布 $p_q(x)$ ，在非重叠区域 $S_{oq} \setminus S_{ou}$ 服从均匀分布，即

$$p_M(x) \begin{cases} \rightarrow p_q(x), & x \in S_{ou} \cap S_{oq} \\ = (1 - P_{Mq}) / (|S_{oq}| - |S_{ou} \cap S_{oq}|), & x \in S_{oq} \setminus S_{ou} \end{cases} \quad (4.4)$$

利用 KL 散度衡量真实分布与模型假设分布之间的差异有[45]：

$$\begin{aligned} D_{KL}(p_q(x) \parallel p_M(x)) &= \sum_{x \in S_{oq}} p_q(x) \log \frac{p_q(x)}{p_M(x)} \\ &= \sum_{x \in S_{ou} \cap S_{oq}} p_q(x) \log \frac{p_q(x)}{p_M(x)} + \sum_{x \in S_{oq} \setminus S_{ou}} p_q(x) \log \frac{p_q(x)}{p_M(x)} \end{aligned} \quad (4.5)$$

根据 (4.4)，在重叠区域

$$\sum_{x \in S_{ou} \cap S_{oq}} p_q(x) \log \frac{p_q(x)}{p_M(x)} \rightarrow \sum_{x \in S_{ou} \cap S_{oq}} p_q(x) \log \frac{p_q(x)}{p_q(x)} = 0 \quad (4.6)$$

在未重叠区域，真实分布 $p_q(x)$ 可能任意，同样根据 (4.4) 有

$$\begin{aligned} \sum_{x \in S_{oq} \setminus S_{ou}} p_q(x) \log \frac{p_q(x)}{p_M(x)} &= \sum_{x \in S_{oq} \setminus S_{ou}} p_q(x) \log p_q(x) - \sum_{x \in S_{oq} \setminus S_{ou}} p_q(x) \log p_M(x) \\ &= -H(p_q) - \sum_{x \in S_{oq} \setminus S_{ou}} p_q(x) \log((1 - P_{Mq}) / (|S_{oq}| - |S_{ou} \cap S_{oq}|)) \\ &= -H(p_q) - \log((1 - P_{Mq}) / (|S_{oq}| - |S_{ou} \cap S_{oq}|)) \sum_{x \in S_{oq} \setminus S_{ou}} p_q(x) \\ &= -H(p_q) - (1 - P_{Mq}) \log((1 - P_{Mq}) / (|S_{oq}| - |S_{ou} \cap S_{oq}|)) \\ &= (1 - P_{Mq}) \log(|S_{oq}| - |S_{ou} \cap S_{oq}|) / (1 - P_{Mq}) - H(p_q) \quad (4.7) \end{aligned}$$

因为 $p_M(x)$ 和 $p_q(x)$ 是定义于同一可测空间的概率分布，我们可以利用总变差距离 (Total Variation Distance, TVD) 和 KL 散度的关系估计模型 M 的泛化误差。根据 Pinsker 不等式有[46]:

$$\begin{aligned} \mathcal{E}_{\text{gen}} &\leq \text{TVD}(p_q, p_M) \leq \sqrt{\frac{1}{2} D_{KL}(p_q(x) \parallel p_M(x))} \\ &= \sqrt{\frac{1}{2} \left(\sum_{x \in S_{ou} \cap S_{oq}} p_q(x) \log \frac{p_q(x)}{p_M(x)} + \sum_{x \in S_{oq} \setminus S_{ou}} p_q(x) \log \frac{p_q(x)}{p_M(x)} \right)} \\ &\approx \sqrt{\frac{1}{2} ((1 - P_{Mq}) \log(|S_{oq}| - |S_{ou} \cap S_{oq}|) / (1 - P_{Mq})) - H(p_q)} \quad (4.8) \end{aligned}$$

由此 (4.3) 式的下半部分得证。因为我们已经假定 $Iu = \langle ou, Tu_h, S_{ou}, M, Tu_m, S_M \rangle$ 和 $Iq = \langle oq, Tq_h, S_{oq}, cq, Tq_m, S_{cq} \rangle$ 皆为可还原信息，状态 $S_M(M, Tu_m)$ 和 $S_{cq}(cq, Tq_m)$ 分别为 S_{ou} 和 S_{oq} 在模型 M 及其接口上的物理实现，因此 (4.3) 式能够表征模型 M 的泛化误差上界。定理证毕。

不同于 VC 维、Rademacher 复杂度等传统泛化理论，定理 5 基于 MLT-MF，以反映训练数据与模型原有信息重叠情况的 P_{Mq} 和 $|S_{oq}| - |S_{ou} \cap S_{oq}|$ ，以及训练数据在非重叠区域的熵 $H(p_q)$ 作为衡量指标，给出了模型泛化误差的显式量化估计，为数据采集与模型设计提供了新的视角和实用指导。特别地，定理 5 揭示了只要问题空间包含于模型的信息空间，就能在统计意义上使模型的泛化误差接近于 0。这是关于机器学习尤为重要的重要规律：

推论 2 (理想模型的泛化误差) 若模型信息包含问题信息，则针对这些问题的泛化误差几乎为 0。

5 结论与展望

本文构建了“机器学习理论元框架 (MLT-MF)”，通过形式化信息模型与六元组定义，统一刻画了机器学习全生命周期的信息流动与因果机制。核心贡献体现在理论整合方面，首次将数据初始化、训练优化与知识迁移纳入统一框架，弥补了传统研究在时间因果链建模上的碎片化缺陷；关于模型可解释性，通过高阶逻辑系统显式表达本体与载

体状态，通过信息可还原性将模型决策映射回到形式化语义空间，确保逻辑透明性，为可解释性提供了严格数学基础；关于伦理安全性，通过嵌入伦理约束谓词，形式化验证模型输出的合规性，为高可信 AI 系统设计提供了判断准则和方法指引；关于泛化误差量化，基于状态覆盖度与 KL 散度，推导获得泛化误差上界公式，揭示了“问题空间需包含于模型知识空间”的核心规律，为训练数据采集和模型优化提供了显式定量分析工具。

尽管取得重要突破，但当前框架假设理想条件下的信息无噪声使能映射，主要针对静态场景中的模型学习和处理逻辑，同时还未涉及多模态、多智能体系统跨本体空间的信息融合与冲突消解。未来仍需从以下方向继续深化：需要扩展至噪声条件下在线学习等复杂动态场景，引入噪声本体与增量覆盖机制；针对多模态、多智能体系统，研究跨本体空间的信息融合与冲突消解，构建分布式 MLT-MF；推动 MLT-MF 与量子计算、生物信息学等领域的交叉，探索量子态的形式化映射和基因调控网络的可解释建模；设计自动化工具链，支持从形式化规则到代码转换，降低 MLT-MF 的工程落地门槛。

致谢

本文撰写和修改中，得到了上海交通大学计算机学院王瑞副教授很多富有卓见的参考意见；在与邱思源、李淳、徐虎和李泽言等四位博士研究生的定期学术讨论中，也获得了很多启发和帮助，在此一并向他们致以诚挚的谢意！

参考文献：

- [1] Rob Ashmore, Radu Calinescu, and Colin Paterson. (2021). Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *ACM Comput. Surv.* 54, 5, Article 111 (June 2022), 39 pages. <https://doi.org/10.1145/3453444>.
- [2] Marius Schlegel and Kai-Uwe Sattler. (2023). Management of Machine Learning Lifecycle Artifacts: A Survey. *SIGMOD Rec.* 51, 4 (December 2022), 18–35. <https://doi.org/10.1145/3582302.3582306>.
- [3] Meng, Z., Zhao, P., Yu, Y., & King, I. (2023). A unified view of deep learning for reaction and retrosynthesis prediction: Current status and future challenges. *arXiv preprint arXiv:2306.15890*.
- [4] Alquier, P. (2024). User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2), 174–303.
- [5] Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial intelligence review*, 18, 77–95.
- [6] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57.
- [7] Glanois, C., Weng, P., Zimmer, M., Li, D., Yang, T., Hao, J., & Liu, W. (2024). A survey on interpretable reinforcement learning. *Machine Learning*, 113(8), 5847–5890.
- [8] Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3), 54–60.
- [9] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612–634.

- [10] Zhang, Y., Hu, S., Zhang, L. Y., Shi, J., Li, M., Liu, X., ... & Jin, H. (2024, May). Why does little robustness help? a further step towards understanding adversarial transferability. In 2024 IEEE Symposium on Security and Privacy (SP) (pp. 3365–3384). IEEE.
- [11] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.
- [12] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- [13] Venkata Pavan Saish, N., Jayashree, J., & Vijayashree, J. (2025). Mathematical Foundations and Applications of Generative AI Models. *Generative Artificial Intelligence (AI) Approaches for Industrial Applications*, 19–45.
- [14] Lee, M. (2023). A mathematical interpretation of autoregressive generative pre-trained transformer and self-supervised learning. *Mathematics*, 11(11), 2451.
- [15] Alvarez Melis, D., & Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.
- [16] Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810.
- [17] Xu, X., Huang, S. L., Zheng, L., & Wornell, G. W. (2022). An information theoretic interpretation to deep neural networks. *Entropy*, 24(1), 135.
- [18] Xu, Jianfeng, Xuefeng Ma, Yanli Shen, Jun Tang, Bin Xu, and Yongjie Qiao (2014). Objective information theory: A Sextuple model and 9 kinds of metrics. 2014 Science and information conference. IEEE, pp. 793 – 802. <https://doi.org/10.1109/SAI.2014.6918277>.
- [19] 许建峰, 汤俊, 马雪峰, 等. (2015). 客观信息的模型和度量研究. *中国科学: 信息科学*[J], 45(3): 336 – 353.
- [20] Jianfeng, X. U., Shuliang, W. A. N. G., Zhenyu, L. I. U., & Yingfei, W. A. N. G. (2021). Objective information theory exemplified in air traffic control system. *Chinese Journal of Electronics*, 30(4), 743–751.
- [21] Jianfeng Xu; Zhenyu Liu; Shuliang Wang; Tao Zheng; Yashi Wang; Yingfei Wang; Yingxu Dang (2023). Foundations and applications of information systems dynamics, *Engineering*, 2023.8, 2023(27): 254–265
- [22] Xu, Jianfeng (2024). Research and Application of General Information Measures Based on a Unified Model. *IEEE Transactions on Computers*. <https://doi.org/10.1109/TC.2024.3349650>.
- [23] Xu, J., Liu, C., Tan, X., Zhu, X., Wu, A., Wan, H., ... & Wu, F. (2025). General Information Metrics for Improving AI Model Training Efficiency. arXiv preprint arXiv:2501.02004.
- [24] C. E. Shannon, “The mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.
- [25] R. Landauer, “The physical nature of information,” *Phys. Lett. A*, vol. 217, nos. 4–5, pp. 188–193, 1996.
- [26] Wiener, Norbert (2019). *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press.

- [27] Hamilton, A. G. (1988). *Logic for mathematicians*. Cambridge University Press.
- [28] 邱思源, (2025). 主观可客观对象状态的形式化表达研究[D]. 上海交通大学.
- [29] 《数学辞海》编辑委员会, 数学辞海, 第五卷, 中国科学技术出版社, 2002.8
- [30] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- [31] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5), 206-215.
- [32] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1-38.
- [33] Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI magazine*, 36(4), 105-114.
- [34] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9), 389-399.
- [35] Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., ... & de Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10).
- [36] Saheb, T., & Saheb, T. (2024). Mapping ethical artificial intelligence policy landscape: a mixed method analysis. *Science and engineering ethics*, 30(2), 9.
- [37] Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28, 689-707.
- [38] Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and machines*, 30(1), 99-120.
- [39] Pant, A., Hoda, R., Tantithamthavorn, C., & Turhan, B. (2024). Ethics in AI through the practitioner’s view: a grounded theory literature review. *Empirical Software Engineering*, 29(3), 67.
- [40] Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5), 988-999.
- [41] Bartlett, P. L., & Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov), 463-482.
- [42] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.
- [43] Advani, M. S., Saxe, A. M., & Sompolsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132, 428-446.
- [44] Perlaza, S. M., & Zou, X. (2024). The generalization error of machine learning algorithms. *arXiv preprint arXiv:2411.12030*.
- [45] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.
- [46] Pinsker, M. S. (1964). Information and information stability of random variables and processes. Holden-Day.

(通讯作者：许建峰 E-mail:xujf@sjtu.edu.cn)